# Prediction of Functional Sites Based on the Fuzzy Oil Drop Model

Michał Bryliński[1,2], Katarzyna Prymula[1,2], Wiktor Jurkowski[1], Marek Kochańczyk[1,3], Ewa Stawowczyk[1], Leszek Konieczny[4], Irena Roterman[1,3*]

**1** Department of Bioinformatics and Telemedicine, Jagiellonian University–Collegium Medicum, Kraków, Poland, **2** Faculty of Chemistry, Jagiellonian University, Kraków, Poland, **3** Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, Kraków, Poland, **4** Institute of Medical Biochemistry, Jagiellonian University–Collegium Medicum, Kraków, Poland

**A description of many biological processes requires knowledge of the 3-D structure of proteins and, in particular, the defined active site responsible for biological function. Many proteins, the genes of which have been identified as the result of human genome sequencing, and which were synthesized experimentally, await identification of their biological activity. Currently used methods do not always yield satisfactory results, and new algorithms need to be developed to recognize the localization of active sites in proteins. This paper describes a computational model that can be used to identify potential areas that are able to interact with other molecules (ligands, substrates, inhibitors, etc.). The model for active site recognition is based on the analysis of hydrophobicity distribution in protein molecules. It is shown, based on the analyses of proteins with known biological activity and of proteins of unknown function, that the region of significantly irregular hydrophobicity distribution in proteins appears to be function related.**

Citation: Bryliński M, Prymula K, Jurkowski W, Kochańczyk M, Stawowczyk E, et al. (2007) Prediction of functional sites based on the Fuzzy Oil Drop model. PLoS Comput Biol 3(5): e94. doi:10.1371/journal.pcbi.0030094

## Introduction

Because of the growing number of structural genomics projects oriented toward obtaining a large number of protein structures in rapid and automated processes [1–4], there is a need to predict protein function (or its functionally important residues) by examining its structure. There have been a variety of efforts in this direction. Some of the techniques used to identify functionally important residues from sequence or structure are based on searching for homologue proteins of known functions [5–8]. However, homologues, particularly when the sequence identity is below 25%, need not have related activities [9–11]. Geometry-based methods have shown that the location of active site residues can be identified by searching for cavities in the protein structure [12] or by docking small molecules onto the structure [13]. The cave localization in silico has been presented on the basis of the characteristics of the normal created for each surface piece [14]. The complex analysis of protein interfaces and their characteristics versus highly divergent areas is presented by Jimenez [15]. Several experimental studies have shown that mutation of residues involved in forming interfaces with other proteins or ligands can also be replaced to produce more stable, but inactive proteins [16–19]. On this basis, several effective algorithms were developed [20,21]. Finally, structural analysis coupled with measures of surface hydrophobicity have been used to identify sites on the surfaces of proteins involved in protein–protein interactions [22].

The Fuzzy Oil Drop (FOD) model presented in this paper is based on an external hydrophobic force field [23–27]. The role of hydrophobic interactions in protein folding [28–31] as well as in protein structure stabilization [32–36] has been known since the classic oil drop model of representing the hydrophobic core in proteins was introduced by Kauzmann [37]. According to this model, the hydrophobic residues tend to be placed in the central part of the protein molecule and in hydrophilic residues on the protein's surface [38–40]. Even the recognition of native versus nonnative protein structures can be to some extent differentiated on the basis of spatial distribution of amino acid hydrophobicity [41–43]. The importance of hydrophobicity distribution has been emphasized, particularly for Rosetta development, when the description of the hydrophobic core significantly increased the performance of the Rosetta program [44]. The discrete system of ellipsoidal centroids was introduced to estimate the concentration of hydrophobic residues, in particular protein zones [44]. The nonrandom hydrophobicity distribution has been proven by Irbäck et al. [45]. However, it was suggested that the core region is not well described by a spheroid of buried residues surrounded by surface residues due to hydrophobic channels that permeate the molecule [46,47]. The FOD model was initially used to simulate the hydrophobic collapse of partially folded proteins. Those structural forms were assumed to represent the early stages of folding (in silico); that model is presented elsewhere [48–50]. The comparison of structures received by folding simulations with their native forms revealed, however, some unexpected results. In the case of native structures, the idealized hydrophobicity distribution satisfying the oil drop–like hydro-

**Editor:** Philip E. Bourne, University of California San Diego, United States of America

**Received** August 14, 2006; **Accepted** April 11, 2007; **Published** May 25, 2007

A previous version of this article appeared as an Early Online Release on April 12, 2007 (doi:10.1371/journal.pcbi.0030094.eor).



**Abbreviations:** CSA, Catalytic Site Atlas; FOD, Fuzzy Oil Drop

* To whom correspondence should be addressed. E-mail: myroterm@cyf-kr.edu.pl

## Author Summary

We present here a method of defining functional site recognition in proteins. The active site (enzymatic cavity or ligand-binding site) is localized on the basis of hydrophobicity deficiency, which is understood as the difference between empirical (dependent on amino acid positions) and idealized (3-D Gauss function, or Fuzzy Oil Drop model) distribution of hydrophobicity. It is assumed that the localization of amino acids representing a high difference of hydrophobic density reveals the functional site. The analysis of the structure of 33 proteins of known biological activity and of 33 proteins of unknown function (with comparable polypeptide chain lengths) seems to verify the applicability of the method to binding cavity localization. The comparative analysis with other methods oriented on biological function is also presented. The validation of predictability accuracy is shown with respect to the enzyme classes.


phobicity partitioning compared with the empirically observed hydrophobicity differs in a specific manner. The high discrepancies between observed and theoretical hydrophobicities within FOD are observed in the area of the binding site [23–26]. It can even be generalized that the location of hydrophobicity differences seems to represent an aim-oriented discrepancy. This simple observation gave us the opportunity to develop a method that was able to recognize functional sites or residues in a protein structure.

In this study, the FOD model is applied to 33 proteins of known function and 33 proteins of unknown function that resulted from structural genomics projects.

## Materials and Methods

### Data

The 33 proteins of known biological activity (Table 1) were selected to verify the reliability of the method. Most of these proteins are enzymes that have well-defined biological function and are deposited in the Catalytic Site Atlas (http://www.ebi.ac.uk/thornton-srv/databases/CSA), a database of templates representing different catalytic mechanisms [51]. The residues identified in this database as active site were used as the criteria to verify the results. Two proteins of known function—rat annexin V, and ButF, the vitamin $B_{12}$-binding protein, which take part in regulation [52] and transport processes [53], respectively—are also included in the test probe.

Reports from structural genomics projects [1–4] have described the solution of a number of proteins with unknown functions. The procedure for potential functional site

**Table 1.** Proteins of Known Function Taken to Analysis

| Organism | Molecule Name | PDB ID | Number | Figures |
|---|---|---|---|---|
| *Aquifex pyrophilus* | Glutamate racemase | 1B73 | 254 (252) | S1, S5 |
| *Bacillus stearothermophilus* | Tyrosyl transfer RNA synthetase | 2TS1 | 419(317) | S2, S7 |
| | Alanine racemase | 1BD0 | 388 (381) | S2, S7 |
| Bacteriophage t4 | Lysozyme | 206L | 164 (162) | S2, S7 |
| *Bos taurus* | Ribonuclease A | 1RBN | 124 | S2, S6 |
| | Carboxypeptidase A | 5CPA | 307 | S2, S7 |
| *Candida albicans* | Phosphomannose isomerase | 1PMI | 440 | 1A, 3A |
| *Equuus caballus* | Alcohol dehydrogenase | 1QLH | 374 | S2, S6 |
| *Escherichia coli* | Methylenetetrahydrofolate reductase | 1B5T | 275 | S1, S5 |
| | Superoxide dismutase | 1ESO | 154 | S1, S6 |
| | Asparagine synthetase | 12AS | 330 (327) | S2, S7 |
| | Methylmalonyl coa decarboxylase | 1EF8 | 261 (256) | S1, S5 |
| | Deoxyribose-phosphate aldolase | 1P1X | 260 (250) | S2, S6 |
| | Endonuclease III | 2ABK | 211 | S2, S6 |
| | Vitamin $B_{12}$ transport protein | 1N2Z | 245 | S1, S6 |
| *Gallus gallus* | Triosophosphate isomerase | 1TPH | 247 (245) | 1B, 3B |
| *Homo sapiens* | Deoxyguanosine kinase | 1JAG | 241 (229) | S1, S6 |
| | Myeloperoxidase | 1MHL | 108 (104) | S1, S6 |
| | Dihydrofolate reductase | 1DHF | 186 (182) | S1, S5 |
| | Protein disulfide isomerase | 1MEK | 120 | 1C, 3C |
| Human immunodeficiency virus | HIV-1 protease | 1A30 | 99 | S1, S5 |
| *Limulus polyphemus* | Arginine kinase | 1BG0 | 356 | S1, S5 |
| *Rattus norvegicus* | Heme oxygenase-1 | 1DVE | 267 (214) | S1, S5 |
| | Annexin V | 1A8A | 318 | S1, S5 |
| | Guanine nucleotide-binding protein | 1BH2 | 315 | S1, S5 |
| *Rhizopus niveus* | Ribonuclease rh | 1BOL | 222 | S1, S5 |
| *Salmo salar* | Trypsin | 1A0J | 223 | S1, S5 |
| *Schizosaccharomyces pombe* | Riboflavin synthase | 1KZL | 208 (202) | S1, S6 |
| *Spinacia oleracea* | Glycolate oxidase | 1GOX | 370 (351) | S1, S6 |
| *Staphylococcus aureus* | 7,8-dihydroneopterin aldolase | 2DHN | 121 | 1D, 3D |
| *Homo sapiens* | Carbonic anhydrase | 1AM6 | 259 (258) | S1, S5 |
| *Nicotiana glutinosa* | Ribonuclease NT | 1VD1 | 217 (203) | S2, S6 |
| *Rhodococcus erythropolis* | Nitrile hydratase | 2AHJ | 206 (192) | S2, S6 |

Number denotes the length of polypeptide chain. The number in parentheses is the number of residues in the polypeptide chain available in the PDB. The last column (on the right) presents the numbers of figures representing results concerning that particular molecule.

doi:10.1371/journal.pcbi.0030094.t001

**Table 2.** Proteins of Unknown Function Taken to Analysis

| Research Group | Organism | PDB ID | Number | Figures |
|---|---|---|---|---|
| **Riken Structural Genomics/Proteomics Initiative (RSGI)** | *Thermus thermophilus* | 2CV9 | 252 | S3, S8 |
| | | 2CVB | 188 (187) | S3, S8 |
| | | 2CW4 | 124 | S3, S8 |
| | | 2CW5 | 255 (235) | S3, S8 |
| | | 2CWY | 94 | S3, S8 |
| | | 2CX0 | 187 (184) | S3, S8 |
| | | 2CXF | 190 (167) | S3, S8 |
| | | 2CXL | 190 (158) | S3, S8 |
| | | 2D4R | 147 (146) | 2C, 4C |
| **Midwest Center for Structural Genomics (MCSG)** | *Pseudomonas aeruginosa* | 2AZP | 318 | 2A, 4A |
| | *Thermotoga maritima* | 2ESH | 118 (114) | S3, S8 |
| | *Helicobacter pylori* | 2EVV | 207 (181) | S3, S8 |
| | *Strptococcus pyogenes* | 2EWC | 126 (120) | S3, S9 |
| | *Bacteroides thetaiotaomicron* | 2F06 | 144 | S4, S9 |
| | *Nitrosomonas europaea* | 2FBL | 153 (144) | S4, S9 |
| **Northeast Structural Genomics Consortium (NESG)** | *Acinetobacter sp.* | 2EW0 | 192 (175) | S3, S9 |
| | *Salmonella enterica* | 2F9C | 334 (320) | S4, S9 |
| | *Bacillus subtilis* | 2FFG | 87 (80) | S4, S10 |
| | *Pseudomonas putida* | 2FFI | 288 (273) | S4, S10 |
| | *Staphylococcus aureus* | 2FFM | 91 (83) | 2D, 4D |
| **Structural Genomics Consortium (SGC)** | *Toxoplasma gondii* | 2F4Z | 193 (145) | S4, S9 |
| | *Homo sapiens* | 2FBM | 291 (251) | S4, S10 |
| | *Plasmodium berghei* | 2FDS | 352 (318) | S4, S10 |
| **Joint Center for Structural Genomics (JCSG)** | *Thermotoga maritima* | 2EWR | 170 (156) | 2B, 4B |
| | *Thermotoga maritima* | 2F4L | 297 (275) | S3, S9 |
| | *Bacillus halodurans* | 2F22 | 144 (142) | SS, S9 |
| **New York Structural Genomics Research Consortium (NYSGRC)** | *Pseudomonas aeruginosa* | 2EUI | 153 | S3, S8 |
| | *Methanococcus jannaschii* | 2F4N | 173 (137) | S4, S9 |
| **Center for Eucaryotic Structural Genomics (CESG)** | *Danio rerio* | 2FB7 | 95 (80) | S4, S9 |
| **Ontario Centre for Structural Proteomics (OCSP)** | *Escherichia coli* | 2F09 | 102 (82) | S4, S9 |
| **Southeast Collabolatory for Structural Genomics (SECSG)** | *Pyrococcus furiosus* | 2F40 | 96 (74) | S4, S9 |
| **Bunker RD, Baker EN, Arcus VL** | *Pyrobaculum aerophilum* | 2FE1 | 156 (130) | S4, S10 |
| **Kang SJ, Park SJ, Jung SJ, Lee BJ** | *Helicobacter pylori* | 1ZHC | 76 | S3, S8 |

Number denotes the length of polypeptide chain. The number in parentheses in the Number column is the number of residues in polypeptide chain available in the PDB. The last column (right) presents numbers of figures representing results concerning that particular molecule.
doi:10.1371/journal.pcbi.0030094.t002

recognition presented in this paper was performed with a set of 33 such proteins deposited in the Protein Data Bank (PDB) (Table 2).

The multimeric proteins were represented solely by their first chain in the PDB file. All molecular visualizations were created with Pymol software [54].

## Hydrophobic Force Field

The FOD hydrophobic force field is based on the assumption that the theoretical hydrophobicity distribution in proteins is represented by the 3-D Gaussian function. The value of this function in a particular $j$-th point within the space occupied by a protein represents the hydrophobicity density at this point:

$$\tilde{H}t_j = \frac{1}{\tilde{H}t_{sum}} \exp\left(\frac{-(x_j-\bar{x})^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_j-\bar{y})^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_j-\bar{z})^2}{2\sigma_z^2}\right) \quad (1)$$

Where $\tilde{H}t_j$ is the theoretical (expected) hydrophobicity of the $j$-th point, $\sigma_x$, $\sigma_y$, $\sigma_z$ are the standard deviations, which depend on the length of polypeptide under consideration [23–26] and the point $(\bar{x},\bar{y},\bar{z})$ is localized in the center of coordinate system (0,0,0) of the highest theoretical hydrophobicity. This simplifies Equation 1:

$$\tilde{H}t_j = \frac{1}{\tilde{H}t_{sum}} \exp\left(\frac{-(x_j)^2}{2\sigma_x^2}\right) \exp\left(\frac{-(y_j)^2}{2\sigma_y^2}\right) \exp\left(\frac{-(z_j)^2}{2\sigma_z^2}\right). \quad (2)$$

The molecule is oriented according to the following procedure: the longest distance between two effective atoms determines the $z$-axis, and the longest distance between projections on the $x$–$y$ plane determines the $x$-axis.

For this orientation of molecules in the coordinate system, the values of $\sigma_x$, $\sigma_y$, $\sigma_z$ parameters are calculated as one-third of the highest $x$, $y$, or $z$ coordinates of the effective atom increased by 9 Å (cutoff distance for hydrophobic interaction) in each direction. The values of the Gaussian function are standardized to give a value of 1.0.

The second component of this force field is an observed (empirical) hydrophobicity distribution formed by the side chains of a protein molecule, and can be expressed using the original function introduced by Levitt [55]. The $j$-th point collects hydrophobicity $\tilde{H}o_j$ as follows:

$$\tilde{H}o_j = \frac{1}{\tilde{H}o_{sum}} \sum_{i=1}^{N} \tilde{H}_i^r \begin{cases} \left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right] \text{ for } r_{ij} \le c \\ \text{otherwise } 0 \end{cases} \quad (3)$$

where $\tilde{H}o_j$ denotes the empirical hydrophobicity value characteristic for the $j$-th point, $N$ is the number of residues in a protein, $\tilde{H}_i^r$ represents the hydrophobicity characteristic for the $i$-th amino acid, $r_{ij}$ is the distance between the $j$-th point and the geometrical center of the $i$-th residue, and $c$ expresses the cutoff distance, which has a fixed value of 9.0 Å, following the original paper [55]. The observed hydrophobicity distribution$\tilde{H}o$ is also standardized. More details concerning the FOD force field are given in recently published papers [23–27].

The similarity of the FOD-based hydrophobic scale with others commonly used for calculations (e.g., the Eisenberg [56] or Doolittle [40] scales) has been shown and discussed in [57]. The differences between these scales seem to be negligible with respect to the problem under consideration. Use of these scales does not change the $\tilde{H}o$ distribution significantly (Equation 3) [57]. The introduction of the FOD-based hydrophobic scale unifies the system for proteins (amino acids) and molecules interacting with proteins, creating stable complexes (ligands).

**Scoring Function.** Since both theoretical $\tilde{H}t$ and observed $\tilde{H}o$ distributions of hydrophobicity are standardized to 1.0 and were calculated for the same set of points (geometrical centers of all residues in a protein), the comparison of these two characteristics is possible. The difference between theoretical and empirical distributions $\Delta\tilde{H}$ expresses the irregularity of hydrophobic core construction. For the $i$-th residue, $\Delta\tilde{H}_i$ is calculated as follows:

$$\Delta\tilde{H}_i = \tilde{H}t_i - \tilde{H}o_i \quad (4)$$

where $\tilde{H}t_i$ and $\tilde{H}o_i$ are the theoretical and observed values of hydrophobicity for the geometric center of the $i$-th residue, respectively.

The maxima of $\Delta\tilde{H}$ recognize the residues representing the hydrophobicity deficiency, which points out the structural irregularity, usually in a function-related area.

**Comparative Analysis.** The SuMo and ProFunc methods (both available on the Web, see urls below) were selected to perform the comparative analysis as to functional site recognition.

*SuMo.* SuMo is a Web tool [58] (http://sumo-pbil.ibcp.fr/cgi-bin/sumo-welcome) that predicts the function of proteins based on the chemistry of the bound ligand. Each ligand and macromolecule part is divided into sets of arbitrary predefined chemical groups. The active site is recognized by a comparison of a minimum of three chemical groups in both compared molecules. SuMo produces a list of probable active sites on default ranked by the number of SuMo groups involved in each given prediction. The active site is described by a set of amino acids and corresponding chemical groups [59].

*ProFunc.* ProFunc [60] is a Web server (http://www.ebi.ac.uk/thornton-srv/databases/ProFunc) devoted to predicting the function of proteins of known 3-D structure and unknown function. The server provides both sequence- and structure-based methods, which may be used in the analysis of proteins. From the group of structure-based methods available on the server, the "reverse templates" 3-D template–based method [61] was chosen and applied to validate the method presented in this study. According to the reverse-template method, the structure itself is broken up into a large number of templates (each containing three residues) that are scanned against a representative set of structures in the PDB [61]. All the hits obtained are scored and ranked. Other homology/sequence-based tools were not taken into account; only methods of similar (structure-based) methodologies were included.

The coordinates of all protein structures under study were submitted to the server in PDB format. The top reverse template–matching structures of known and unknown functions were used in our comparative analysis.

### Result Verification

The residues annotated in CSA as those playing roles in catalytic activity were used as the gold standard to verify the reliability of the results received according to the FOD model.

To indicate the most meaningful amino acids considered by the FOD model to be located in the functional site, the calculation of percentiles was used to identify the threshold for selection of $\Delta\tilde{H}$ maxima, which are distinguished as belonging to the functional site. It is possible to do so, because the quantitative results expressing the level of $\Delta\tilde{H}$ can be taken as the criteria for discrimination. For a set of measurements arranged in order of magnitude, the *p-th* percentile is the value that has $p$ percent of the measurements below it and $(100 - p)$ percent above it. In this analysis, the 95th percentile was used. In other words, among the analyzed data, 95% of values were below the 95th percentile threshold, and only the 5% above the threshold was taken into consideration.

The same validation method cannot be used in the SuMo or ProFunc methods because of their different types of output data. They produce only the numbers of amino acids that potentially belong to functional sites and total scores (based on which given set of amino acid residues is assessed and what functional site is proposed). This is why the percentage of commonly classified residues was calculated for each protein molecule by taking the best hit by ProFunc (according to the score value) and the solution most relevant to the FOD-based results by SuMo.

## Results

### Functional Site Recognition in Proteins of Differentiated Biological Activity

The proteins of known biological activity (Table 1) and protein structures of unknown function that resulted from structural genomics projects (Table 2) were examined for the locations of their functional sites. Table 3 summarizes the results of the method application and comparison with experimental observations (CSA classification). The first column presents the protein under consideration and the list of residues recognized by CSA. For two proteins (rat annexin V and ButF), residues that are in direct contact with ligand [62,63] and/or are part of the functional site are given [64].

**Table 3.** Biological Activity–Related Residues as Recognized in Proteins of Known Biological Function Using Methods Discussed in This Paper

| PDB ID/Number of AAs, Name of Active Site | Fuzzy Oil Drop Model | | SuMo Method | | ProFunc Method | |
|---|---|---|---|---|---|---|
| | Number of AAs | AA Names | PDB ID/Number of AAs | FOD/SuMo Ratio | PDB ID/Number of AAs | Score |
| **1A30**/25 ASP, 26 THR | 25, 28, **29**, **30**, **31** | ASP, ALA, **ASP**, **ASP**, **THR** | 1STC/**29**, **30** | 0.4 | 2F80/**31**, 75, 86 | 870.31 |
| **206L**/11 GLU, 20 ASP | **11**, **30**, 12, *105*, 10, 145, **104** | **GLU**, GLY, GLY, ***GLN***, ASP, ARG, **PHE** | 1OWZ/**30**, **104** | 0.5 | 176L/88 TYR, 100 ILE, 101 ASN | 860 |
| **1RBN**/12 HIS, 120 PHE, 119 HIS, 41 LYS | **12**, **120**, *11*, **45**, **119**, 7, 41 | **HIS**, **PHE**, GLN, **THR**, **HIS**, LYS, LYS | 1RPG/***11***, **12**, **45**, **120** | 0.22 | 1IZR/**12 HIS**, ***44 ASN***, 47 VAL | 840 |
| **1BH2**/43 GLU, 178 ARG, 204 GLN, 181 THR | *180*, **43**, **178**, ***44***, **45**, 150, 270 | *LYS*, **GLU**, **ARG**, **SER**, **GLY**, ASP, LYS | 1BH2/**43**, ***44***, **45**, **178** | 0.2 | 1BH2/29 SER, 30 PRO, 209 TRP | 940 |
| **1KZL**/97 HIS, 146 SER, 48 CYS, 185 ASP, 41 SER | 183, 144, 141, ***186***, **185**, 140, 184, 104, 189, 133, 132 | GLU, GLY, ALA, **GLN**, **ASP**, ILE, VAL, ASP, LYS, TYR, LYS | 2ACV/**185**, ***186*** | 0.2 | 1KZL/45 ASN, *47 THR*, 73 LEU | 1,020 |
| **1BG0**/126 ARG, 280 ARG, 22 ARG, 225 GLU, 309 ARG | **126**, **280**, 274, 225, **314**, **330**, 226, 273 | **ARG**, **ARG**, ASN, GLU, **GLU**, **ARG**, ASP, THR | 1SD0/**126**, **280**, **314** | 0.25 | 1M15/129 ARG, 278 THR, **330 ARG** | 1,000 |
| **1JAG**/142 ARG, 70 GLU | **142**, **70**, **141**, **118**, 51, 44 | ***ARG***, **GLU**, GLU, **ARG**, LYS, GLU | 1BXR/**118**, ***141***, **142** | 0.4 | 1JAG/***143 SER***, 229 TRP, 242 LEU | 980 |
| **1MHL**/91 GLN, 95 HIS | *94*, 98, 100, 101, 29, 28, 31 | *ASP*, ASP, THR, PRO, PHE, ALA, ARG | NO HIT | | 1D2V/9 THR, 14 CYS, 20 PRO | 480 |
| **1A0J**/57 HIS, 102 ASP, 195 SER | 102, **195**, **196**, 213, 229, 43, **193**, 194, 44, 53, 198, 31, 32, 139, 30 | ASP, **SER**, **GLY**, VAL, THR, GLY, GLY, ASP, GLY, VAL, PRO, ALA, SER, SER, GLN | 1GJB/**193**, **195** | 0.285 | 1A0J/55 ALA, ***196 GLY***, ***197 GLY*** | 1,040 |
| **1DVE**/139 GLY, 136 ARG, 140 ASP, 135 THR, 143 GLY, 25 HIS, 58 TYR | **139**, **136**, 140, 137, 58, 85, 62 | **GLY**, **ARG**, ASP, TYR, TYR, ARG, GLU | 1J2C/**136**, **139** | 0.16 | 1J02/74 TYR, 129 LEU, 132 HIS | 920 |
| **1B5T**/28 GLU, 120 ASP | 28, 120, **275**, **183** | GLU, ASP, **TYR**, **GLN** | 1SLY/**183**, **275** | 0.18 | 1B5T/***27 PHE***, 283 SER, 287 CYS | 960 |
| **1B73**/70 CYS, 7 ASP, 8 SER, 178 CYS | ***70***, 7, 8, *69*, 11, 71, 72, **178**, 40, 114, ***177***, 117, *179*, 200 | ***CYS***, ASP, SER, *ALA*, GLY, ASN, THR, **CYS**, GLY, THR, ***GLY***, THR, *THR*, SER | 1ZUW/**70**, ***71***, ***177*** | 0.31 | 1B74/111 VAL, 123 TYR, ***177 GLY*** | 960 |
| **1BOL**/105 GLU, 109 HIS, 46 HIS | **105**, 47, **46**, 48, 30, 137, 136, 200 | **GLU**, GLY, **HIS**, LEU, ASN, ALA, LYS, ASP | 1UCD/**46**, **105** | 0.22 | 1BOL/109 HIS, 110 GLY, 133 TYR | 940 |
| **2TSI**/82 LYS, 230 LYS, 233 LYS, 86 ARG | 230, 47, 80, **194**, **78**, 48, 193, **192**, **38**, **195**, 221, 50, 51, 269, 173, 10 | LYS,GLY, SER, **ASP**, **ASP**, HIS, SER, **GLY**, **ASP**, **GLN**, PRO, ALA, THR, TYR, GLN, ARG | 1JIL/**38**, **78**, **173**, **192**, **194**, **195** | 0.27 | 2TSI/34 TYR, 186 CYS, 189 GLN | 960 |
| **1GOX**/254 HIS, 157 ASP, 257 ARG, 129 TYR | 230, *255*, 77, 78, **285**, 76, 309, **286**, 308, 38, **287**, **288**, **289**, 39, 290, 293 | LYS, *GLY*, PRO, THR, **ASP**, ALA, ARG, **GLY**, GLY, ASN, **GLY**, **VAL**, **ARG**, ARG, ARG, ARG | 1GOX/**285**, **286**, **287**, **288**, **289** | 0.55 | 1GOX/81 GLN, 92 THR, 95 ALA | 1,080 |
| **1DHF**/22 LEU, 30 GLU | 115, 35, 56, **38**, 55, 132, **48**, ***122*** | VAL, GLN, THR, **THR**, LYS, LYS, **ASN**, ***LYS*** | 1JX2/**38**, **48** | 0.18 | 1KMV/8 VAL, ***121 TYR***, 136 THR | 840 |
| **1ESO**/61 HIS | **44**, 42, 43, 72, 83, 84, 140, 40, 122 | **HIS**, GLY, PHE, PRO, PRO, ALA, GLU, GLU, ASP | 1BZO/**44** | 0.09 | 1ESO/46 HIS, 61 HIS, 118 HIS | 960 |
| **5CPA**/71 ARG, 127 ARG, 270 GLU | ***72***, **69**, **127**, **196**, 68, 67, 144, **112**, 194, **111**, 65, **108**, 175 | ***GLU***, **HIS**, **ARG**, **HIS**, ILE, GLY, ASN, **ASN**, SER, **THR**, ASP, GLU, GLU | 1ARM/**69**, ***72***, **127**, **196** | 0.4 | 5CPA/**108 GLU**, **111 THR**, **112 ASN** | 1,000 |
| **1QLH**/48 SER, 51 HIS | **48**, ***47***, **46**, **174**, 68, 175, 369, 203, 292, 178, 173, 319, 177, 179, 91, 92, 206, **324** | **SER**, ***ARG***, **CYS**, **CYS**, GLU, GLY, ARG, VAL, VAL, THR, GLY, PHE, SER, GLY, PRO, LEU, SER, **SER** | 1HTB/**46**, **48**, **174**, **324** | 0.31 | 1N8K/145 THR, 150 THR, 151 VAL | 1,140 |
| **12AS**/100 ARG, 46 ASP, 116 GLN | 72, 71, **100**, 114, **46**, **116**, 74, **294**, ***115***, 45, 97, 47, **251**, 214, 248 | SER, HIS, **ARG**, VAL, **ASP**, **GLN**, ALA, **GLY**, ***ASP***, GLN, LYS, ASN, **SER**, ARG, GLU | 12AS/**100**, **116**, **251** | 0.75 | 12AS/119 TRP, 121 ARG, **294 GLY** | 980 |
| **1BD0**/311 CYS, 39 LYS, 265 TYR | 355, 357, 40, **354**, 353, **43**, 342, 39, 343, 341, 243 | GLU, PRO, ALA, **TYR**, ASN, **TYR**, ILE, LYS, ASP, SER, GLU | 1BD0/**43**, **354** | 0.1 | 1BD0/293 GLY, 288 TRP, 332 ILE | 1,000 |
| **1EF8**/110 GLY, 66 HIS, 140 TYR | 113, 114, 89, 147, 117, **93** | GLU, MET, THR, ASN, SER, **GLN** | 1JDF/**93** | 0.17 | 1EF8/101 SER, 103 VAL, 127 SER | 1,000 |
| **1MEK**/37 GLY, 39 CYS, 36 CYS, 38 HIS | **64**, 86, 24, 25, 26, 27, 59, ***61***, 87, 89, 90, 97, 115 | **LYS**, LYS, HIS, LYS, TYR, LEU, GLU, ***ARG***, PHE, ARG, ASN, LYS, ARG | NO HIT | | 1MEK/46 TYR, ***62 LEU***, **64 LYS** | 860 |
| **1P1X**/167 LYS, 201 LYS, 102 ASP | 167, **201**, 137, 102, **16** | LYS, **LYS**, LYS, ASP, **ASP** | NO HIT | | 1P1X/**16 ASP**, 47 CYS, **201 LYS** | 1,040 |
| **2ABK**/138 ASP, 120 LYS | 138, 120, 136, 180, 119, 39, 184, 44 | ASP, LYS, ALA, ILE, ARG, SER, ARG, ASP | NO HIT | | 2ABK/144 VAL, 148 THR, 182 HIS | 900 |
| **2DHN**/22 GLU, 100 LYS | 97, 32, 115, 34, ***95*** | ARG, ASP, GLU, THR, ***LYS*** | NO HIT | | 1DHN/84 ILE, ***94 THR***, 118 ARG | 900 |
| **1PMI**/111 GLN, 304 ARG, 294 GLU | **111**, **285**, 287, **138**, 279, 294, **284**, 283, 102, 140, 100, 277, 48 | **GLN**, **HIS**, TYR, **GLU**, LEU, GLU, **PRO**, ASP, LEU, ALA, LYS, MET, GLU | 1PMI/**111**, **138**, **284**, **285** | 0.22 | 1PMI/206 PHE, 261 CYS, 323 TYR | 980 |
| **1TPH**/11 ASN, 95 HIS, 13 LYS, 165 GLU, 171 GLY | 97, **165**, 64, 98, 65, 11, **95**, 13 | ASN, **HIS**, LYS, GLU, **GLU**, GLN, ARG, ASN | 1TPB/**95**, **165** | 0.15 | 1SW0/***164 TYR***, 208 TYR, 226 VAL | 980 |
| **1AM6**/199 THR, 106 GLU | **199**, **106**, **96**, **200**, 117, **119**, **94**, 92, 67 | **THR**, **GLU**, **HIS**, **THR**, GLU, **HIS**, **HIS**, GLN, ASN | 1BNV/**94**, **96**, **106**, **119**, **199**, **200** | 0.43 | 1LUG/30 PRO, ***107 HIS***, 209 TRP | 1,000 |

**Table 3.** Continued.

| PDB ID/Number of AAs, Name of Active Site | Fuzzy Oil Drop Model | | SuMo Method | | ProFunc Method | |
|---|---|---|---|---|---|---|
| | Number of AAs | AA Names | PDB ID/Number of AAs | FOD/SuMo Ratio | PDB ID/Number of AAs | Score |
| **1VD1**/93 GLU, 97 HIS, 39 HIS | **93**, **39**, *96*, *10*, 12, 11, 171, 37, 158 | **GLU**, **HIS**, *LYS*, VAL, GLN, GLN, GLN, SER, GLU | 1VD1/**39**, **93** | 0.18 | 1VD3/*9 PHE*, 119 SER, 175 CYS | 920 |
| **2AHJ**/113 SER | **113**, 128, 115, 123, 133, 90 | **SER**, LYS, THR, PRO, ARG, GLN | 2AHJ/**113** | 0.08 | 2CZ1/56 GLY, 60 VAL, 64 TRP | 880 |
| **1A8A/Calcium-binding site;** 28 GLY, 30 GLY, 31 THR, 70 GLU, 100 GLY, 102 GLY, 103 THR, 142 ASP, 181 GLY, 182 GLU, 183 LEU, 184 LYS,185 TRP, 186 GLY, 187 THR, 259 GLY, 261 GLY, 262 THR, 301 ASP; **ion channel**: 112 GLU, 271 ARG, 98 ASP, 117 ARG, 17 GLU, 78 GLU, 95 GLU | 86, 89, 90, 93, 106, **110**, ***111***, 114, 115, 119, 269, 273, 274 | SER, TYR, ASP, GLU, LYS, ***GLU***, **ILE**, SER, ARG, GLU, ARG, SER, ARG | 2RAN/***111*** | 0.08 | 2RAN/***109 THR***, ***111 ILE***, 147 TYR | 880 |
| **1N2Z**/30 SER, 31 PRO, 32 ALA, 50 TYR, 85 TRP, 87 GLY, 196 TRP, 241 SER, 242 ASP, 245 GLU, 246 ARG | 32, 108, 110, 174, 176, *245*, **246**, *247* | ALA, ASP, THR, SER, GLN, *GLU*, **ARG**, *ALA* | 1N2Z/**246** | 0.04 | 1N2Z/28 THR, 29 LEU, 48 SER | 960 |

Column 1 residues are recognized as active site (CSA database) in the order according to increased distance versus the ligand position (geometrical center of ligand or the averaged position of amino acids responsible for biological function as found in literature).
Bold numbers are given for residues recognized by FOD and at least one of two analyzed methods (SuMo, ProFunc).
Underlined numbers represent residues found to be function-related (according to CSA database).
Italicized numbers are shown when the position of the amino acids differ by 1 (closest neighbors) versus the CSA classification or versus the position found by FOD.
For FOD versus SuMo comparison, amino acids common for both methods are pointed out with the PDB protein code, which functional site was found to be related with the protein under consideration.
FOD/SuMo ratio expresses the part of amino acids common for FOD versus the total number of residues pointed out by SuMo.
ProFunc score values are given according to the program classification.
doi:10.1371/journal.pcbi.0030094.t003

**Table 4.** Biological Activity–Related Residues as Recognized in Proteins of Unknown Biological Function Using Methods Discussed in This Paper

| Fuzzy Oil Drop Model | | SuMo Method | | | ProFunc Method | |
|---|---|---|---|---|---|---|
| PDB ID/Number of AAs | AA Name | PDB ID/ Number of AAs | FOD/SuMo Ratio | Score | PDB ID/ Number of AAs | Score |
| **1ZHC/30,** 53, 57 | **ASP,** LYS, LYS | 1KQS/**30** | 0.25 | 2.99/3.027 | 1ZHC/ 13 LYS, 18 HIS, 61 HIS | 680 |
| **2AZP**/14, **15, 16,** 17, 57, 60, 61, **80,** 81 | GLU, **PRO, THR,** ARG, ASP, VAL, GLY, **ASN,** ASN | 4KBP/**15, 16, 80** | 0.75 | 2.1/3.15 | 2AZP/92 THR, 94 GLY, 98 SER | 1,140 |
| **2CV9/*8*, 35, 37,** 65, **111, 143, 170** | *ASP*, **ASN, GLU,** ASN, **GLN, GLU, HIS** | 1H2G/**35, 37, 111, 143** | 0.8 | 2.999/3.75 | 2CVB/***7 GLY, 9 VAL,*** 194 THR; 1T70/168 GLY, 169 THR, **170 HIS** | 1,100 |
| **2CVB/6,** 118, 119, **132,** 133, **167,** 169, 170 | **GLU,** PRO, GLU, **HIS,** GLY, **GLU,** PRO, ALA | 1OA1/**6, 132, 167** | 0.6 | 2.7/3.35 | 2CVB/42 CYS, 72 ASN, 103 ASP | 1,020 |
| **2CW4/31, 54, 55,** 56, 57 | **GLN,** GLU, ASN, LEU, LYS | 1TSY/**54, 55** | 0.67 | 2.1/2.77 | 2CVL/60 LEU, 66 GLY, 119 CYS; 2B33/10 PRO, **31 GLN,** 116 GLU | 880 |
| **2CW5**/44, 46, 47, 48, 49, **50,** 51, **54** | ASP, ARG, ARG, ALA, ALA, **TYR,** ALA, **GLU** | 1ZI9/**50, 54** | 1 | 2.43/3.45 | 2CW5/189 THR, 232 GLU, 119 CYS | 1,000 |
| **2CWY/41, *42,*** 45, 84 | *GLY*, *VAL*, LEU, LEU | 1BQQ/**41** | 0.33 | 2.43/3.0 | 2CWY/5 TRP, 39 LEU, ***40 GLN*** | 920 |
| **2CX0**/77, **78,** 91, **141** | PRO, **THR,** LYS, **VAL** | 1U0F/**141** | 0.25 | 3.60/5.50 | 2CX1/66 CYS, **78 THR,** 168 HIS | 900 |
| **2CXF**/152, 165, **166, 167,** 169, **170, *183,*** 186, 187, 242 | GLU, THR, **ALA, SER,** LYS, **ASP, *ALA,*** ARG, LEU, MET | 1YI7/**166, 167, 170** | 0.75 | 2.1/3.3 | 2CXF/154 VAL, ***184 TRP,*** 197 TYR | 920 |
| **2CXL/183, 184, 186,** 187, 188, 191, 193 | **ALA, TRP, ARG,** LEU, ALA, GLN, LYS | 1N2C/**183, 184, 186** | 0.75 | 2.1/3.05 | 2DWK/154 VAL, **184 TRP,** 197 TYR | 920 |
| **2D4R**/49, **51,** 64, **66, 87,** 91, **130** | SER, **TRP,** GLU, **GLU, TYR,** TRP, **ASN** | 1NSI/**51, 66, 87, 130** | 0.67 | 2.4/3.6 | 2D4R/31 LEU, **51 TRP, 66 GLU** | 920 |
| **2ESH**/8, 9, 11, **12, 96, 100,** 103 | GLY, PHE, GLY, **TRP, SER, MET,** ARG | 1LU2/**12, 96, 100** | 0.6 | 2.15/3.03 | 2D4R/31 LEU, 51 TRP, 66 GLU | 920 |
| **2EUI/65, 82, 83** | ***GLN,* ASN, ASP** | 1TKK/**65, 82, 83** | 0.75 | 2.1/3.43 | 2EUI/51 LEU, 62 GLY, ***64 CYS*** | 860 |
| **2EVV**/55, 56, **72, 73,** 95 | GLU, LEU, **TRP, VAL,** GLN | 1JX6/**72, 73** | 0.66 | 2.1/3.22 | 2EVV/41 ILE, **72 TRP,** 87 SER | 980 |
| **2EW0/38, 39,** 40 | **GLN, GLY,** ILE | 1VRQ/**38, 39** | 0.67 | 2.1/4.2 | 2EW0/25 THR, 27 ILE, 28 TYR | 980 |
| **2EWC**/27, **28, 29, *115*** | LEU, **ASN, TYR, *ASP*** | 1CF4/**28, 29** | 0.67 | 2.15/3.15 | 2EWC/49 MET, 50 GLU, ***116 GLY*** | 694 |
| **2EWR/*27,* 45,** 47, 91, 98 | ***THR,* ASP,** GLN, GLU, LYS | 1SBE/**27, 45** | 0.67 | 2.43/3.9 | 2FCL/***28 GLY,*** 31 SER, 130 LEU | 800 |
| **2F06/5, 46, 69, 99, *112*** | **GLN, ARG, THR, TYR, *VAL*** | 1K56/**5, 46, 69, 99** | 1 | 2.73/3.6 | 2FO6/73 GLY, ***113 ILE,*** 121 CYS | 880 |
| **2F09/5,** 13, 14, **16, 17, *18,*** 23, 58, 59, 61, 80 | **THR,** LYS, PRO, **THR, VAL, *LYS,*** ARG, LYS, GLY, GLU, PRO | 2BME/**5, 16, 17** | 0.6 | 2.42/2.75 | 2F09/**5 THR, *19 LEU,*** 48 TYR | 840 |
| **2F22**/42, **43,** 45, **46, *48*** | HIS, **ILE,** ARG, **VAL, *ASP*** | 1DQO/**43, 46** | 0.67 | 2.1/3.67 | 2F22/***47 ARG,*** 83 LEU, 115 GLU | 940 |
| **2F40/5, 53, *57,* 64** | **LYS, GLU, *LYS,* GLU** | 1S76/**5, 64** | 0.4 | 2.39/3.0 | 2F40/44 ILE, **53 GLU, *56 GLU*** | 880 |
| **2F4L/148, 149, 150,** 151, **264,** 274 | **ASP, THR, LYS,** GLU, **SER,** LYS | 1TLL/**148, 149, 150, 264** | 1 | 2.43/4.94 | 2F4L/124 GLY, 126 ILE, 175 HIS | 1,100 |
| **2F4N**/61, 63, 64, **65, 114** | THR, GLY, ALA, **TYR, TYR** | 2A99/65, 114 | 1 | 2.45/3.35 | 2F4N/106 GLY, 144 TYR, 145 ALA | 960 |
| **2F4Z**/70, 96, 111, 112, 144, 151 | GLU, LEU, MET, LYS, ILE, ILE | NO HIT | — | — | 2F4Z/89 TYR, 119 HIS, 187 THR | 940 |
| **2F9C**/119, 136, **137,** 138, 163, **164,** 165, 183 | SER, SER, **GLU,** ILE, SER, **ARG,** ILE, GLU | 1VJJ/**137, 164** | 1 | 2.2/3.83 | 2F9C/59 CYS, 60 TRP, 61 ILE | 960 |
| **2FB7**/21, 23, ***33*** | LYS, SER, ***GLU*** | 1S6J/**33** | 0.25 | 2.87/2.87 | 2FB7/***32 TYR,*** 77 PHE, 79 GLY | 700 |
| **2FBL**/7, 25, **84, 98, *113,* 143** | LYS, GLN, **LYS, GLU, *GLU,* ASN** | 1AXD/**84, 143** | 0.67 | 2.1/4.2 | 2FBL/**98 GLN, *112 VAL,*** 130 TRP | 820 |
| **2FBM**/299, 306, 334, 396, **397, 398, 425** | LEU, LYS, PHE, ALA, **SER, ILE, ASP** | 1EXM/**397, 398, 425** | 1 | 2.43/4.23 | 2FBM/394 LEU, ***426 GLY,*** 512 GLU | 1,000 |
| 2FDS/5, 9, **97, 104, 134, 135,** 288 | LYS, ARG, **LYS, LYS, PRO, THR,** GLU | 1S5D/**97, 134, 135** | 1 | 2.1/3.3 | 2FDS/103 TYR, **104 LYS,** 106 ASN | 960 |
| **2FE1/6, 7, 8,** 42 | **ASP, ALA, SER,** GLU | 1NT4/**6, 7, 8** | 0.75 | 2.45/2.77 | 2FE1/35 THR, 103 TYR, 114 LEU | 1,000 |
| **2FFG/39, 40, *49*** | **VAL, LYS, *GLU*** | 5EAT/**39, 40** | 0.67 | 2.1/2.7 | 2FFG/37 CYS, ***50 LEU,*** 73 GLU | 720 |
| **2FFI/17, 18,** 110, **141, 166,** 197, 236 | **SER, HIS,** ARG, **GLU, ASP,** LYS, ASP | 1BH5/**18, 141, 166** | 1 | 2.4/11.04 | 2FFI/**17 SER, 18 HIS,** 58 VAL | 1,060 |
| **2FFM**/39, **44,** 47, **52,** 65, 77 | PRO, **ASP,** LYS, **LYS,** LYS, LYS | 1UEJ/**44, 52** | 0.67 | 2.15/3.03 | 2FFM/62 SER, 63 VAL, 71 TRP | 780 |

Bold numbers are given for residues recognized by FOD and at least one of two analyzed methods (SuMo, ProFunc).
Italicized numbers indicate when the positions of amino acids differ by 1 (closest neighbors) versus the CSA classification or versus the position found by FOD.
SuMo results are characterized by: numbers of amino acid common for FOD and SuMo.
FOD/SuMo ratio expresses the part of amino acids common for FOD versus the total number of residues pointed out by SuMo and additionally by the relation between the SuMo score of the solution closest to the one based on the FOD model (highest number of common positions) and the score value of best hit, as estimated by SuMo.
ProFunc score values are given according to the program classification.
doi:10.1371/journal.pcbi.0030094.t004

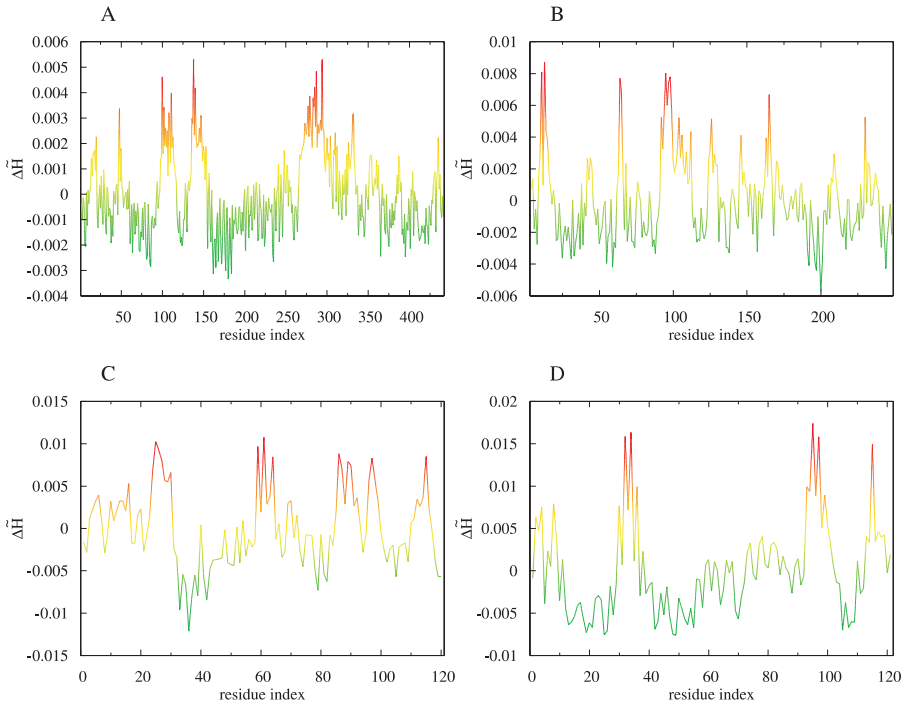


**Figure 1.** Profile Plots of Hydrophobicity Deviation $\Delta\tilde{H}$ per Amino Acid Obtained for Exemplary Proteins of Known Function

(A) Phosphomannose isomerase and (B) triosophosphate isomerase are examples of the high agreement with experimental data.
(C) Protein disulfide isomerase and (D) 7,8-dihydroneopterin aldolase are examples of low agreement.
The common color scale is introduced: red, high $\Delta\tilde{H}$; yellow, middle $\Delta\tilde{H}$; green, low and negative $\Delta\tilde{H}$.
doi:10.1371/journal.pcbi.0030094.g001

In Table 3, the columns representing FOD results show the numbers of residues recognized by this method: agreement with CSA classification (underlined), and residues defined by two methods—FOD and at least one of two other methods (SuMo, ProFunc) as biological activity-related residues (in bold). Where the position of the amino acids differed by 1 (closest neighbors) versus the CSA classification or versus the position found by SuMo or ProFunc, the numbers are in italics in Table 3. The description of the SuMo and ProFunc columns in Table 3 is given below (Comparative Analysis).

The residues recognized as potentially responsible for binding site creation in proteins of unrecognized biological function are given in Table 4.

Profile plots of $\Delta\tilde{H}$ were used to identify the positions recognized by the FOD model as related to functional sites. The profile plots of $\Delta\tilde{H}$ were examined for proteins of known and unknown biological activity (Figures 1, S1, and S2; and Figures 2, S3, and S4; respectively). The residues with the highest $\Delta\tilde{H}$ appeared as peaks in the profile plots and were predicted to be functionally important. The values of $\Delta\tilde{H}$ indicate the level of hydrophobicity irregularity. It is interpreted that the higher the $\Delta\tilde{H}$value, the higher the deficiency of hydrophobicity with respect to its idealized distribution according to Gauss function. Thus, the $\Delta\tilde{H}$ maxima identified as being represented by a particular amino acid point out the residues in the surrounding area where the hydrophobicity deficiency is significant. In most cases, this deficiency is caused simply by the presence of a cavity or by the highly irregular distribution of side chains. The $\Delta\tilde{H}$ profile together with the color scale visualizes the magnitude of the irregularity. The same scale applied to the 3-D presentation of the protein molecule is able to visualize the location of high $\Delta\tilde{H}$ values, particularly in the protein structure. It can be seen that the residues with high $\Delta\tilde{H}$ values appear to be placed in close mutual vicinity, often creating a cleft, which can be responsible for ligand (substrate) binding.

The 3-D representations for selected proteins of known function are shown in Figure 3, and for selected proteins of unknown biological function in Figure 4. Other proteins under consideration are presented in Figures S5–S7 and Figures S8–S10.

The color scale expressing the magnitude of $\Delta\tilde{H}$ is as follows: red, high $\Delta\tilde{H}$; yellow, average $\Delta\tilde{H}$; green, low and negative $\Delta\tilde{H}$. The white color denotes the experimentally verified amino acids as responsible for catalytic activity (according to the CSA database). In most cases, the set of amino acids selected according to the FOD model is larger than the set of residues classified by CSA. This is because the $\Delta\tilde{H}$ profile also selects amino acids that are close in space, which create well-defined putative cavities that accompany the residues responsible for enzymatic activity. Amino acids indicated by FOD as belonging to the binding cavity are in space filling form.

The molecules presented in Figure 3A and 3B are selected to show the best results; the molecules presented in Figure 3C and 3D demonstrate the cases of low accordance. Some of the protein molecules with high $\Delta\tilde{H}$ values shown in Figure 3A and 3B appeared to be highly accordant to the active site location. Other proteins with high $\Delta\tilde{H}$ values (Figure 3C and 3D) are not exactly located in the positions of the amino acids that make up the catalytic site. Nevertheless, the analysis of the larger set of proteins may suggest that the specificity of

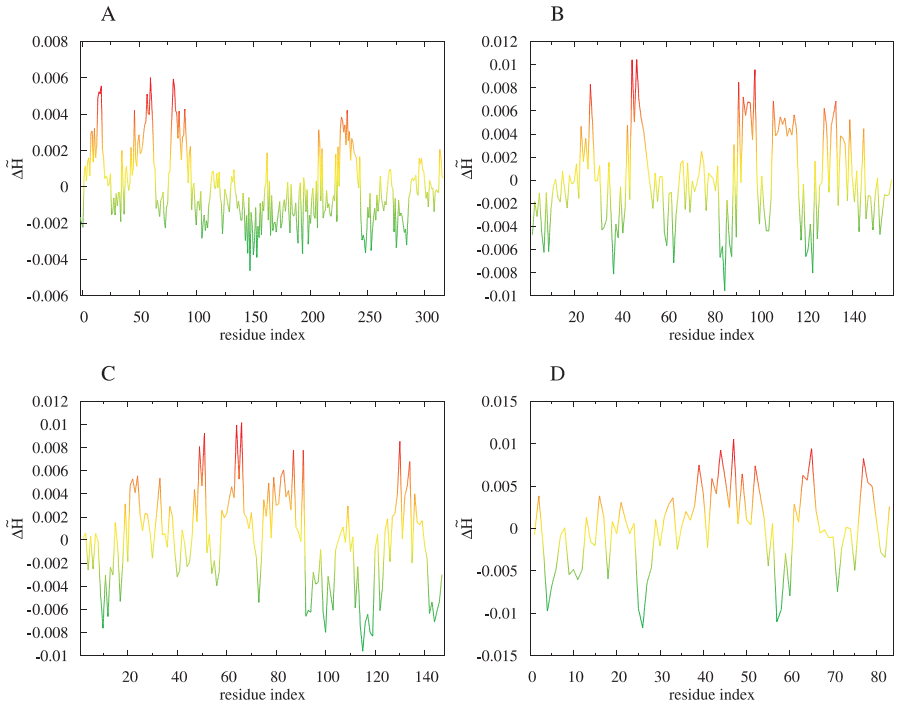


**Figure 2.** Profile Plots of Hydrophobicity Deviation $\Delta\tilde{H}$ per Amino Acid Obtained for Exemplary Proteins of Unknown Function

The protein identified in the genome of *Pseudomonas aeruginosa* (A) and the protein identified in the genome of *Thermotoga maritima* (B) are examples representing close localization of residues of high $\Delta\tilde{H}$. The protein originated in the *Thermus thermophilus* genome (C) and the protein originated the *Staphylococcus aureus* genome (D) are examples of dispersed localization of residues representing high $\Delta\tilde{H}$. The common color scale (same as in Figure 1) is introduced: (low and negative $\Delta\tilde{H}$ proteins need additional analysis of their specificity).

doi:10.1371/journal.pcbi.0030094.g002

the mutual location of the residues represented by high $\Delta\tilde{H}$ values versus the position of the enzymatic site may be classified according to enzyme specificity.

One hypothesis is that the residues responsible for complex fixation (protein and ligand or substrate) were selected by the FOD model. Another explanation for the mismatch between experimentally identified and automatically identified residues is simply that for multimeric chains, only the first chain was present in the analysis.

## Comparative Analysis

The results summarize the comparison of the model applied to identify the ligand-binding site and two other methods dedicated to the same purpose: ProFunc and SuMo are given in Table 3 for proteins of known biological function and in Table 4 for proteins of unknown biological function. Table 3 presents the list of proteins (the PDB accession numbers are given) accompanied by the amino acids identified as function-related according to CSA classification.

SuMo results (for each SuMo search in question) show the comparison with the FOD model for only one example of a functional site found by SuMo and present the residue numbers, which appeared to be common for these two methods (column 4 of Table 3). The limitation to compare only one SuMo result for one search is caused by the specificity of output generated by the SuMo procedure, which produces an enormous number of possible solutions for one particular protein molecule (in most cases, thousands of variants). Each solution is presented with regard to another protein (PDB number given), the functional site of which seems to be related to that found in the molecule under analysis. This procedure proposes a list of functional sites that sometimes represent changed functionality (e.g., ligands of different structure/characteristics are bound). One functional site with a functional site of the same/closest properties is selected. The presentation of all results is impossible to present here in complete form.

In column 5 of Table 3, the ratio of commonly recognized residues to the number of all residues recognized by SuMo for that hit is shown. As we see, the total number of amino acids classified by SuMo in most cases is the same or exceeds the number identified by the FOD model.

The numbers given in the last two columns (ProFunc) of Table 3 represent positions of amino acids recognized by ProFunc by its best hit and method score. This is why the number of commonly recognized residues (given in bold) is lower than in the SuMo comparison.

The results describing the analysis of proteins of unknown biological function are shown in Table 4. The presentation is similar to that for proteins of known biological function with an obvious lack of underlined positions (no CSA classification available). The SuMo results are additionally characterized by the relation between the SuMo score of the solution closest to that based on the FOD model (highest number of common positions) and the score value of best hit, as estimated by SuMo.

The comparison of the methods selected for analysis is generally very difficult. The SuMo and ProFunc methods represent the methodology of the stochastic nature. The FOD seems to be a more heuristic method. SuMo and ProFunc produce very large outputs with long lists of possible

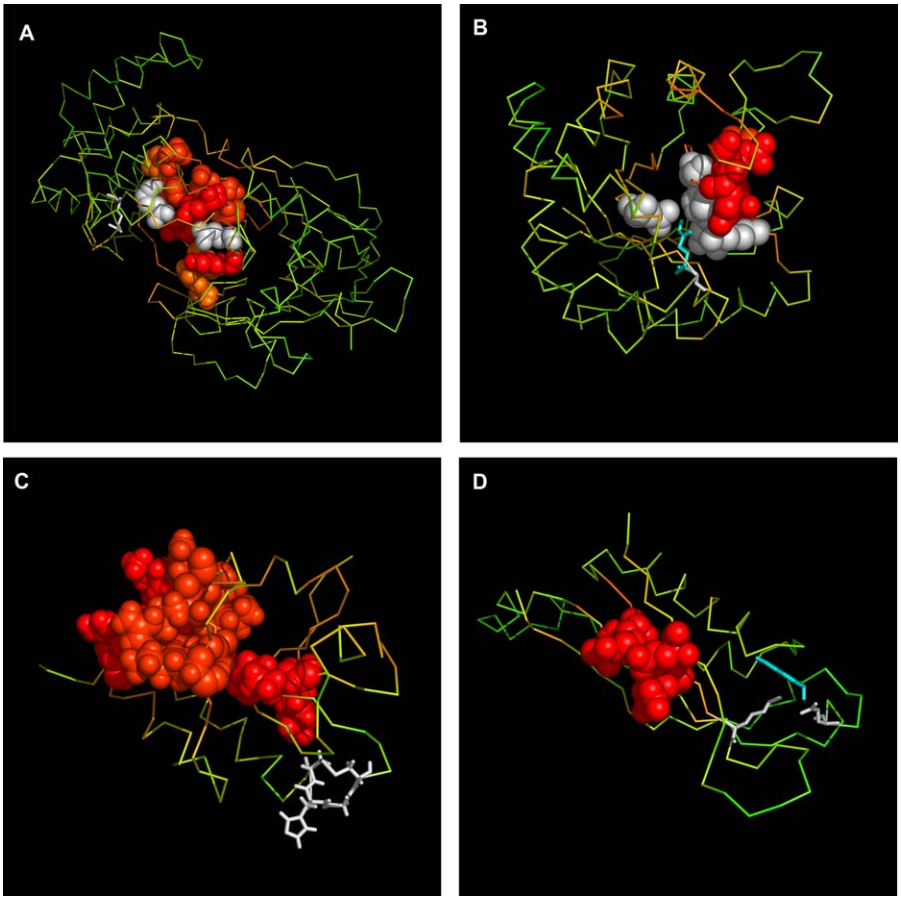


**Figure 3.** The 3-D Representation of Proteins of Known Biological Activity with Binding Site Recognized

Phosphomannose isomerase (A) and triosophosphate isomerase (B) are examples of the high agreement with experimental data. Protein disulfide isomerase (C) and 7,8-dihydroneopterin aldolase (D) are examples of low agreement. Amino acids indicated by FOD as belonging to the binding cavity are in CPK form. The common color scale (same as in Figure 1) is introduced. The white color denotes the experimentally verified amino acids as active site (identification according to the CSA database).

doi:10.1371/journal.pcbi.0030094.g003

approaches. Each of them is characterized by the scoring number calculated according to the number of contacts (pairs of amino acids) responsible for ligand–protein interaction. However, the number of residues commonly recognized by at least two analyzed methods seems to be quite high.

Taking into account a very large discrepancy in the results of one particular method, the level of mutual accordance seems to be satisfactory.

## Result Validation

Tables 5 and 6 present the results aimed toward validating the FOD model–based results. The values present error levels calculated for the methods under consideration. These calculations take into account the number of mismatched residues versus the CSA, SuMo, and ProFunc classifications. Tables 5 and 6 also include comparisons versus functional site amino acids estimated by the $\Delta\tilde{H}$ above the 95th percentile value.

The proteins of known biological function are characterized in Table 5, and the proteins of unknown biological function are characterized in Table 6. The false negative (below diagonal) and false positive (above diagonal) classifications are given as average (for all analyzed proteins) percentages of mismatched residues.

The comparison is expressed by the level of error measured in the percentage of mismatched residues. The left value in each table cell was calculated by taking into account the exact amino acid numbers. The value on the right side expresses the percentage of mismatched residues when the tolerance of $(i + 2)/(i - 2)$ amino acids (the positions of the residues) is taken into consideration.

The FOD results are based on the $\Delta\tilde{H}$ profile along the polypeptide chain. The search for the percentile optimally discriminating the residues belonging to those classified by CSA can be performed. The $\Delta\tilde{H}$ values above the 95th percentile value appeared to be the best approach of local $\Delta\tilde{H}$ maxima as the criteria for function-related residue classification. The results of the comparison of the 95th percentile are shown in the "FOD 95th percentile" column.

The interpretation of values given in Tables 5 and 6 is as follows. For example, in FOD versus ProFunc cases, 86% of residues found by the FOD method were not selected by ProFunc (false positives). Taking the amino acids with $(i+2)/(i-2)$ tolerance, the level decreases to 73%.

In false negative cases, 81% of residues selected by ProFunc were not selected by FOD (65% when closest neighbors were taken into account).

This study is not designed to give a thorough comparison

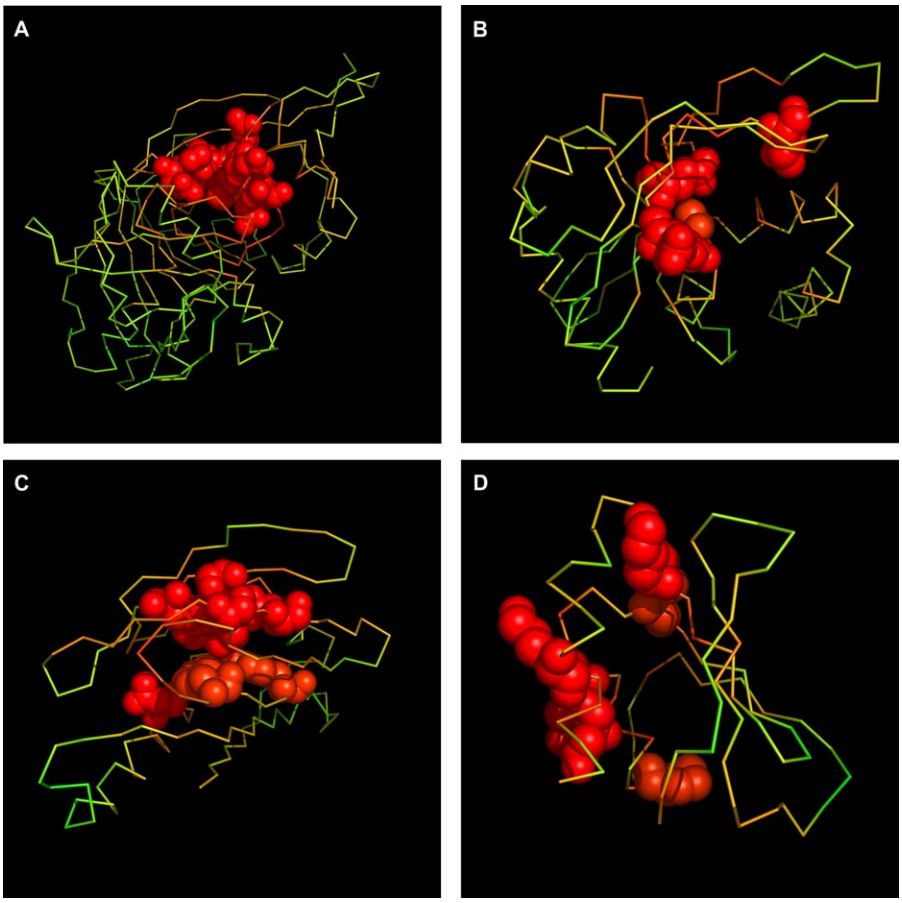


**Figure 4.** The 3-D Representation of Proteins of Unknown Biological Activity with Binding Site Recognized

The protein identified in the *Pseudomonas aeruginosa* genome (A) and the protein identified in the *Thermotoga maritima* genome (B) are examples representing close localization of residues of high $\Delta\tilde{H}$. The protein originated in the *Thermus thermophilus* genome (C) and the protein originated in the *Staphylococcus aureus* genome (D) are examples of dispersed localization of residues representing high $\Delta\tilde{H}$. The common color scale (same as in Figure 1) is introduced.

doi:10.1371/journal.pcbi.0030094.g004

of functional site tools, nor is it meant to review the current advances in this field. Therefore, the mutual comparisons between SuMo and ProFunc, SuMo and CSA, and ProFunc and CSA are not presented here.

Additional analysis summarizing the applicability of the presented method is also shown in Table 7. It is shown that the correctness of the FOD model depends on the enzyme class. Values in Table 7 express the percentages of the residues identified by the FOD method versus those identified by CSA. The highest agreement was found for the EC.3 category (hydrolases), where almost 70% of residues classified by CSA were found by the FOD model. The functional sites in enzymes belonging to the EC.4 (lyases) and EC.6 (ligases) classes were recognized quite well (more than 60%). The lowest agreement was found for the EC.2 class (transferases), where the percentage of correctly predicted amino acids (versus CSA classification) was about 20% (this seems nonrepresentative due to the low number of proteins under consideration in this class).

The specificity of the active sites in particular enzymatic classes will be analyzed in future publications with respect to the FOD methodology. The larger number of proteins belonging to particular enzyme classes will be taken into consideration in the prospective analysis with respect to the applicability of the FOD model as the tool for functional site recognition. The increased number of proteins representing a particular enzyme class may clarify also the applicability of the method in relation to the detailed type of reaction catalyzed.

## Discussion

The recognition of functional sites in protein molecules is important for the identification of biological activity. The fully automatic method is highly expected. In analogy to the methods applied for protein structure prediction, the ligand-binding site can be recognized on the basis of *comparative methods* (according to CASP [critical assessment of structure prediction] classification). The alternate possibility is to search for a ligand-binding site using *new fold* (according to CASP classification) techniques that use only the structure of individual proteins.

The FOD method presented here identifies the potentially function-related amino acids. In contrast to SuMo and ProFunc, which are based on comparative analysis, the FOD method is of heuristic form, taking as its criterion the individual local hydrophobicity deficiency in a particular protein body.

**Table 5.** Error Analysis for Proteins of Known Biological Function

| False Positive/False Negative | Fuzzy Oil Drop | Fuzzy Oil Drop 95th Percentile | Catalytic Site Atlas | SuMo Method | ProFunc Method |
|---|---|---|---|---|---|
| Fuzzy Oil Drop method | — | 55/35 | 80/66 | 70/59 | 86/73 |
| Fuzzy Oil Drop 95th percentile | 58/36 | — | 54/33 | — | 87/69 |
| Catalytic Site Atlas | 46/36 | — | — | — | — |
| SuMo method | 71 | — | — | — | — |
| ProFunc method | 81/65 | 91/78 | — | — | — |

doi:10.1371/journal.pcbi.0030094.t005

The ligands' (as cofactors or cosubstrates) presence makes the biological activity possible for some proteins. The enzymatic activity also requires substrate binding. The presence in the cavity of high specificity versus ligand/substrate is needed for this kind of interaction. The location of the cavity (dependent on the protein character) in protein molecules seems to be well recognized by the FOD model.

The part of the protein molecule with high hydrophobic deficiency is recognized as a possible ligand-binding site (or active site). Some results received according to the FOD model seem to be quite satisfactory (Figure 1A and 1B and Figure 3A and 3B). The catalytic mechanisms of enzymes are quite differentiated and require appropriate molecular structures. The analysis of their specificity may clarify the origin of failure (Figure 1C and 1D and Figure 3C and 3D). The possible protein–protein complex creation (not taken into consideration in this analysis) may significantly influence the results (e.g., Figures S1 and S6). Two proteins (in Figures 1C and 3C, and in Figures 1C, 3C, S2, and S7) of common enzymatic specificity (disulphide isomerase) have been recognized on the basis of the FOD method as highly similar with respect to the mutual orientation of residues involved in cavity creation. The specificity of enzymes with respect to their active site construction is the aim of prospective analysis, which will be published soon, as well as analysis of proteins responsible for biological functions other than enzymatic (e.g., proteins responsible for transport as given in Table 3).

The calcium-binding sites in annexin V are not recognized by FOD, although the ion channel–creating residues are pointed out by this method according to expectations for the method of biological function recognition.

The FOD model may also represent the specific hydrophobic environment for protein folding and was initially aimed at the simulation of the hydrophobic collapse of partially folded proteins. The heuristic model of protein folding, according to which the folding polypeptide is directed to follow the hydrophobicity distribution, is represented by the 3-D Gaussian function. The external force field may direct the folding process toward the hydrophobic core creation. The resulting structure appeared to be dissatisfactory, particularly because of the absence of a ligand-binding site in the final structural form. The presence of a ligand in the folding environment may ensure the specific binding cavity creation. Thus, it seems to be important or even necessary.

The comparative analysis of the results of the FOD-based method with the results of SuMo and ProFunc (Tables 3–6) reveals the very high similarity of obtained results. The methods use different criteria for classification. The exhaustive comparative analysis of the results obtained by the application of different methods seems to be necessary and has been taken into consideration; this will be published soon together with explanation of the source of these differences.

The proteins shown in this paper represent mostly enzymes of varying biological activity, the relation of which to the character of the results will be the object of independent research.

It is generally accepted that globular proteins consist of a hydrophobic core and a hydrophilic surface [36,40]. However, the core region is not well described by a spheroid of hydrophobic residues surrounded by hydrophilic residues due to channels that permeate the molecule [46,47]. The FOD model, when applied to protein structure, characterizes the hydrophobicity density in a continuous form by pointing out the irregularities in a hydrophobic core construction disturbing the regularity of hydrophobicity distribution

**Table 6.** Error Analysis for Proteins of Unknown Biological Function

| False Positive/False Negative | Fuzzy Oil Drop Model | Fuzzy Oil Drop 95th Percentile | SuMo Method | ProFunc Method |
|---|---|---|---|---|
| Fuzzy Oil Drop Model | — | 63/35 | 56/40 | 82/59 |
| Fuzzy Oil Drop 95th Percentile | 68/42 | — | — | 94/72 |
| SuMo Method | 30 | — | — | — |
| ProFunc Method | 87/63 | 93/71 | — | — |

The values measuring the disagreement (error) expressed in percentages for proteins of unknown biological activity. The Fuzzy Oil Drop column takes into consideration the amino acids representing maxima on $\Delta H$ profile and that are localized in the close mutual vicinity. The values in the top row and last column are calculated for false positive results, and those in the first column and last row for false negative results. The values on the left side express the level of error when using exact numbers of amino acids; the values on the right side express the level of error when amino acids on positions +2 and/or −2 versus the exact number are taken into account.

doi:10.1371/journal.pcbi.0030094.t006

**Table 7.** Correctness of the Fuzzy Oil Drop Method as Dependent on the Enzyme Category

| International Classification of Enzymes | Percentage of Residues Recognized by Fuzzy Oil Drop versus the CSA Classification | Number of Enzymes Present in Analysis |
|---|---|---|
| E.1 (oxido-reductases) | 47.3 | 7 |
| E.2 (transferases) | 20 | 1 |
| E.3 (hydrolases) | 69 | 8 |
| E.4 (lyases) | 60 | 5 |
| E.5 (isomerases) | 45 | 6 |
| E.6 (ligases) | 62.5 | 2 |

doi:10.1371/journal.pcbi.0030094.t007

[23–26]. Those irregularities seem to be good markers for ligand-binding sites or functionally important residues.

Methods dedicated to active site recognition have been widely developed: SARIG [65], Q-SITE FINDER [66], HIPPO (SPROUT) [67,68], FEATURE [69–71], THEMATICS [72–74], APROPOS [75], DRUGSITE [76], and LIGSITE [77], to mention just a few. Limitation to two methods (SuMo and ProFunc) for comparative analysis in this paper is due to the very large variability of the models when applied.

The method described in this paper is assumed to be applied for active site identification for a large set of proteins, the structure of which is planned to be generated using different methods (FOD and ROSETTA [78]). The project geared toward biological activity identification in never born proteins (NBPs) is assumed to deliver the molecules of pharmacological application [79,80]. This is the main scientific goal for pharmacology application in the EuChinaGrid project.

The FOD method is available at http://bioinformatics.cm-uj.krakow.pl/activesite.

## Supporting Information

**Figure S1.** Other Proteins Listed in Table 1 Presented as Described in Figure 1

Found at doi:10.1371/journal.pcbi.0030094.sg001 (5.1 MB TIF).

**Figure S2.** Continuation of Proteins Listed in Table 1 Presented as Described in Figure 1

Found at doi:10.1371/journal.pcbi.0030094.sg002 (9.6 MB EPS).

**Figure S3.** Other Proteins Listed in Table 2 Presented as Described in Figure 2

Found at doi:10.1371/journal.pcbi.0030094.sg003 (8.3 MB EPS).

**Figure S4.** Continuation of Proteins Listed in Table 2 Presented as Described in Figure 2

Found at doi:10.1371/journal.pcbi.0030094.sg004 (7.9 MB EPS).

**Figure S5.** Other Proteins Listed in Table 2 Presented as Described in Figure 3

Found at doi:10.1371/journal.pcbi.0030094.sg005 (8.0 MB TIF).

**Figure S6.** Continuation of Other Proteins Listed in Table 2 Presented as Described in Figure 3

Found at doi:10.1371/journal.pcbi.0030094.sg006 (8.0 MB TIF).

**Figure S7.** Continuation of Other Proteins Listed in Table 2 Presented as Described in Figure 3

Found at doi:10.1371/journal.pcbi.0030094.sg007 (2.9 MB TIF).

**Figure S8.** Other Proteins Listed in Table 2 Presented as Described in Figure 4

Found at doi:10.1371/journal.pcbi.0030094.sg008 (7.3 MB TIF).

**Figure S9.** Continuation of Other Proteins Listed in Table 2 Presented as Described in Figure 4

Found at doi:10.1371/journal.pcbi.0030094.sg009 (8.4 MB TIF).

**Figure S10.** Continuation of Other Proteins Listed in Table 2 Presented as Described in Figure 4

Found at doi:10.1371/journal.pcbi.0030094.sg010 (3.5 MB TIF).

### Accession Numbers

The Protein Data Bank (http://www.rcsb.org/pdb) accession numbers for the proteins discussed in this paper are: rat annexin V (1A8A), ButF (1N2Z), phosphomannose isomerase (1PMI), triosophosphate isomerase (1TPH), protein disulfide isomerase (1MEK), 7,8-dihydroneopterin aldolase (2DHN), protein identified in the *Pseudomonas aeruginosa* genome (2AZP), protein identified in the *Thermotoga maritima* genome (2EWR), protein originated in the *Thermus thermophilus* genome (2D4R), protein originated in the *Staphylococcus aureus* genome (2FFM), myeloperoxidase (1MHL), and riboflavin synthase (1KZL).

## Acknowledgments

**Author contributions.** IR conceived and designed the experiments. MB, KP, and WJ performed the experiments. ES and LK analyzed the data. MK contributed reagents/materials/analysis tools. IR wrote the paper.

**Funding.** This research was supported by the Polish State Committee for Scientific Research (KBN) grant 3 T11F 003 28 and Collegium Medicum grants 501/P/133/L and WŁ/222/P/L. This work has been funded by the European Commission EUChinaGRID project (contract 026634).

**Competing interests.** The authors have declared that no competing interests exist.

### References

1. Burley SK, Almo SC, Bonanno JB, Capel M, Chance MR, et al. (1999) Structural genomics: Beyond the human genome project. Nat Genet 23: 151–157.
2. Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, et al. (2002) Structural genomics: A pipeline for providing structures for the biologist. Protein Sci 11: 723–738.
3. Goulding CW, Apostol M, Anderson DH, Gill HS, Smith CV, et al. (2002) The TB structural genomics consortium: Providing a structural foundation for drug discovery. Curr Drug Targets Infect Disord 2: 121–141.
4. Chandonia JM, Earnest TN, Brenner SE (2004) Structural genomics and structural biology: Compare and contrast. Genome Biol 5: 343.
5. Zvelebil MJ, Sternberg MJ (1988) Analysis and prediction of the location of catalytic residues in enzymes. Protein Eng 2: 127–138.
6. Wallace AC, Borkakoti N, Thornton JM (1997) TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 6: 2308–2323.
7. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, et al. (1998) Predicting function: From genes to genomes and back. J Mol Biol 283: 707–725.

8. Skolnick J, Fetrow JS (2000) From genes to protein structure and function: Novel applications of computational approaches in the genomic era. Trends Biotechnol 18: 34–39.
9. Hegyi H, Gerstein M (1999) The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. J Mol Biol 288: 147–164.
10. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41: 98–107.
11. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297: 233–249.
12. Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. Protein Sci 7: 1884–1897.
13. Oshiro CM, Kuntz ID, Dixon JS (1995) Flexible ligand docking using a genetic algorithm. J Comput Aided Mol Des 9: 113–130.
14. Lamb ML, Burdick KW, Toba S, Young MM, Skillman AG, et al. (2001) Design, docking, and evaluation of multiple libraries against multiple targets. Proteins 42: 296–318.
15. Jimenez JL (2005) Does structural and chemical divergence play a role in precluding undesirable protein interactions? Proteins 59: 757–764.
16. Meiering EM, Serrano L, Fersht AR (1992) Effect of active site residues in barnase on activity and stability. J Mol Biol 225: 585–589.
17. Zhang J, Liu ZP, Jones TA, Gierasch LM, Sambrook JF (1992) Mutating the charged residues in the binding pocket of cellular retinoic acid-binding protein simultaneously reduces its binding affinity to retinoic acid and increases its thermostability. Proteins 13: 87–99.
18. Shoichet BK, Baase WA, Kuroki R, Matthews BW (1995) A relationship between protein stability and protein function. Proc Natl Acad Sci U S A 92: 452–456.
19. Kanaya S, Oobatake M, Liu Y (1996) Thermal stability of *Escherichia coli* ribonuclease HI and its active site mutants in the presence and absence of the $Mg^{2+}$ ion. Proposal of a novel catalytic role for Glu48. J Biol Chem 271: 32729–32736.
20. Elcock AH (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 312: 885–896.
21. Ondrechen MJ, Clifton JG, Ringe D (2001) THEMATICS: A simple computational predictor of enzyme function from structure. Proc Natl Acad Sci U S A 98: 12473–12478.
22. Jones S, Thornton JM (1997) Analysis of protein–protein interaction sites using surface patches. J Mol Biol 272: 121–132.
23. Brylinski M, Konieczny L, Roterman I (2006) Ligation site in proteins recognized in silico. Bioinformation 1: 127–129.
24. Brylinski M, Konieczny L, Roterman I (2006) Fuzzy-oil-drop hydrophobic force field—A model to represent late-stage folding (in silico) of lysozyme. J Biomol Struct Dyn 23: 519–528.
25. Brylinski M, Konieczny L, Roterman I (2006) Hydrophobic collapse in (in silico) protein folding. Comp Biol Chem 30: 255–267.
26. Brylinski M, Konieczny L, Roterman I (2006) Hydrophobic collapse in late-stage folding (in silico) of bovine pancreatic trypsin inhibitor. Biochimie 88: 1229–1239.
27. Konieczny L, Brylinski M, Roterman I (2006) Gauss-function-based model of hydrophobicity density in proteins. In Silico Biol 6: 0002.
28. Dill KA (1990) Dominant forces in protein folding. Biochemistry 29: 7133–7155.
29. Pace CN, Shirley BA, McNutt M, Gajiwala K (1996) Forces contributing to the conformational stability of proteins. FASEB J 10: 75–83.
30. Baldwin RL (2002) Making a network of hydrophobic clusters. Science 295: 1657–1658.
31. Finney JL, Bowron DT, Daniel RM, Timmins PA, Roberts MA (2003) Molecular and mesoscale structures in hydrophobically driven aqueous solutions. Biophys Chem 105: 391–409.
32. Klotz IM (1970) Comparison of molecular structures of proteins: Helix content; distribution of apolar residues. Arch Biochem Biophys 138: 704–706.
33. Klapper MH (1971) On the nature of the protein interior. Biochim Biophys Acta 229: 557–566.
34. Meirovitch H, Scheraga HA (1980) Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids. Macromolecules 13: 1398–1405.
35. Meirovitch H, Scheraga HA (1980) Empirical studies of hydrophobicity. 2. Distribution of the hydrophobic, hydrophilic, neutral, and ambivalent amino acids in the interior and exterior layers of native proteins. Macromolecules 13: 1406–1414.
36. Meirovitch H, Scheraga HA (1981) Empirical studies of hydrophobicity. 3. Radial distribution of clusters of hydrophobic and hydrophilic amino acids. Macromolecules 14: 340–345.
37. Kauzmann W (1959) Some factors in the interpretation of protein denaturation. Adv Protein Chem 14: 1–63.
38. Chothia C (1975) Structural invariants in protein folding. Nature 254: 304–308.
39. Rose GD, Roy S (1980) Hydrophobic basis of packing in globular proteins. Proc Natl Acad Sci U S A 77: 4643–4647.
40. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105–132.
41. Novotny J, Rashin AA, Bruccoleri RE (1988) Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4: 19–30.
42. Baumann G, Frommel C, Sander C (1989) Polarity as a criterion in protein design. Protein Eng 2: 329–334.
43. Holm H, Sander C (1992) Evaluation of protein models by atomic solvation preference. J Mol Biol 225: 93–105.
44. Bonneau R, Strauss CE, Baker D (2001) Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins 43: 1–11.
45. Irback A, Peterson C, Potthast F (1996) Evidence for nonrandom hydrophobicity structures in protein chains. Proc Natl Acad Sci U S A 93: 9533–9538.
46. Crippen GM, Kuntz ID (1978) A survey of atom packing in globular proteins. Int J Pept Protein Res 12: 47–56.
47. Kuntz ID, Crippen GM (1979) Protein densities. Int J Pept Protein Res 13: 223–228.
48. Brylinski M, Jurkowski W, Konieczny L, Roterman I (2004) Limited conformational space for early-stage protein folding simulation. Bioinformatics 20: 199–205.
49. Jurkowski W, Brylinski M, Konieczny L, Wiiniowski Z, Roterman I (2004) Conformational subspace in simulation of early-stage protein folding. Proteins 55: 115–127.
50. Brylinski M, Konieczny L, Czerwonko P, Jurkowski W, Roterman I (2005) Early-stage folding in proteins (in silico) sequence-to-structure relation. J Biomed Biotechnol 2: 65–79.
51. Porter CT, Bartlett GJ, Thornton JM (2004) The catalytic site atlas: A resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32: D129–D133.
52. Von der Marka K, Mollenhauerb J (1997) Annexin V interactions with collagen. Cell Mol Life Sci 53: 539–545.
53. Locher KP (2004) Structure and mechanism of ABC transporters. Curr Opin Struct Biol 14: 426–431.
54. DeLano WL (2002) The PyMOL molecular graphics system. Available: http://www.pymol.org. Accessed 14 February 2007.
55. Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104: 59–107.
56. Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol. 179: 125–142.
57. Brylinski M, Konieczny L, Roterman I (2007) Hydrophobic collapse: Late stage folding simulation of human $\alpha$ and $\beta$ hemoglobin chains. Int J Bioinf Res Appl. In press.
58. Jambon M, Andrieu O, Combet C, Deléage G, Delfaud F, et al. (2006), The SuMo server: 3D search for protein functional sites. Bioinformatics 21: 3929–3930.
59. Jambon M, Imberty A, Deléage G, Geourjon C (2003) A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 52: 137–145.
60. Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: A server for predicting protein function from 3D structure. Nucleic Acids Res 33: W89–W93.
61. Laskowski RA, Watson JD, Thornton JM (2005). Protein function prediction using local 3D templates. J Mol Biol 351: 614–626.
62. Huber R, Berendes R, Burger A, Luecke H, Karshikov A (1992) Annexin V-crystal structure and its implications on function. Behring Inst Mitt 91: 107–125.
63. Karpowich NK, Huang HH, Smith PC, Hunt JF (2003) Crystal structures of the BtuF periplasmic-binding protein for vitamin $B_{12}$ suggest a functionally important reduction in protein mobility upon ligand binding. J Biol Chem 278: 8429–8434.
64. Kourie JI, Wood HB (2000) Biophysical and molecular properties of annexin-formed channels. Prog Biophys Mol Biol 73: 91–134.
65. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, et al. (2004) Network analysis of protein structures identifies functional residues. J Mol Biol 344: 1135–1146.
66. Laurie AT, Jackson RM (2005) Q-SiteFinder: An energy-based method for the prediction of protein-ligand binding sites. Bioinformatics. 21: 1908–1916.
67. Gillet VJ, Myatt G, Zsoldos Z, Johnson AP (1995) SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. Perspect Drug Discov Design 3: 34–50.
68. Law JMS, Fung DYK, Zsoldos Z, Simon A, Szabo Z, et al. (2003) Validation of the SPROUT de novo design program. J Mol Struct: THEOCHEM 651–657; 666–667.
69. Wei L, Altman RB (1998) Recognizing protein binding sites using statistical descriptions of their 3D environments. Pac Symp Biocomput 497–508.
70. Liang MP, Banatao DR, Klein TE, Brutlag DL, Altman RB (2003) WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures, Nucleic Acids Res 31: 3324–3327.
71. Banatao DR, Altman RB, Klein TE (2003) Microenvironment analysis and identification of magnesium binding sites in RNA. Nucleic Acids Res 31: 4450–4460.

72. Ko J, Murga LF, Wei Y, Ondrechen MJ (2005) Prediction of active sites for protein structures from computed chemical properties, Bioinformatics 21 (Supplement 1): i258–265.
73. Shehadi IA, Abyzov A, Uzun A, Wei Y, Murga LF, et al. (2005) Active site prediction for comparative model structures with thematics. J Bioinform Comput Biol 3: 127–143.
74. Ko J, Murga LF, Andre P, Yang H, Ondrechen MJ, et al. (2005) Statistical criteria for the identification of protein active sites using theoretical microscopic titration curves. Proteins 59: 183–195.
75. Peters KP, Fauck J, Frommel C (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol 256: 201–213.
76. An J, Totrov M, Abagyan R (2004) Comprehensive identification of druggable protein ligand binding sites. Genome Inform 15: 31–41.
77. Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15: 359–363.
78. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci U S A 103: 5361–5366.
79. Chiarabelli C, Vrijbloed JW, Thomas RM, Luisi PL (2006) Investigation of de novo totally random biosequences. Part I: A general method for in vitro selection of folded domains from a random polypeptide library displayed on phage. Chem Biodivers 3: 827–839.
80. Chiarabelli C, Vrijbloed JW, de Lucrezia D, Thomas RM, Stano P, et al. (2006) Investigation of de novo totally random biosequences, Part II: On the folding frequency in a totally random library of de novo proteins obtained by phage display. Chem Biodivers 3: 840–859.